

# Gradient Coupling Effect of Poisoning Attacks in Federated Learning

Wenqi Wei

Fordham University and Georgia Institute of Technology  
[wenqiwei@fordham.edu](mailto:wenqiwei@fordham.edu)

Ling Liu

Georgia Institute of Technology  
[ling.liu@cc.gatech.edu](mailto:ling.liu@cc.gatech.edu)

## Abstract

*Poisoning Attack is a dominating threat in distributed learning, where the mediator has limited control over the distributed client contributing to the joint model. In this paper, we present a comprehensive study on the coupling effect of poisoning attacks from three perspectives. First, we identify the theoretical foundation of the weak coupling phenomenon of gradient eigenvalues when under the poisoning attack. Second, we analyze the behavior of gradient coupling under four scenarios: adaptive attacker, skewed client selection, Non-IID data distribution, and different gradient window sizes. We study when the weak coupling effect would fail as the attack indicator. Last, we examine the coupling effect by revisiting several existing poisoning mitigation approaches. Through formal analysis and extensive empirical evidence, we show under what conditions the weak coupling effect of poisoning attacks can serve as forensic evidence for attack mitigation in federated learning and how it interacts with the existing defenses.*

**Keywords:** federated learning, poisoning attacks, security analysis

## 1. Introduction

Federated learning [16] enables collaborative and distributed model training for many applications, such as next word prediction [4] and electronic health record understanding [20]. In every round, the central server distributes the current joint model to a random subset of participants. Each client trains with their local data and submits the local model update to the server. Then, the server aggregates these local model updates into the new joint model for the next round of training.

The lack of control over client's data in federated learning and transparency in the clients' updates create the space for exploiting training data manipulation with two attack objectives: (i) denial of service (DoS) [2, 8], which prevents the convergence of the global model or makes the model converge to a bad minimum. (ii) targeted poisoning [1, 21, 23], which assumes a small percentage of malicious clients and is targeted at objects of a specific source class (victim) to malfunction towards a target class while keeping the service quality on the rest of data. The latter is believed more difficult but more motivated since the adversaries can tailor the attack to any adverse goal while remaining under the radar. Real-world targeted poison attacks have shown that federated learning-trained object detectors can be hijacked to misdetect objects of designated classes by changing the class label of objects of a source class, e.g., labeling the stop sign as the speed limit [23]. The attack may also modify the presence of the bounding boxes, e.g., making the detection box for person disappear [7].

Existing poisoning detection strategies have observed that the eigenvalues of the gradient covariance between the poisoned and the benign gradient can be separated. This weak coupling effect enables the removal of the potentially poisoned model updates [12, 22, 24, 26]. A key question is whether one can simply conclude with high confidence that the existence of the weak coupling effect is a sign of poisoning attacks and that the smaller cluster in the separation must reflect the poisoned gradients to be removed. Meanwhile, differential privacy noise [15, 19] and byzantine-robust aggregation [1, 8] are commonly considered for poisoning mitigation. While these approaches lead to the tight coupling of poisoned and benign gradients, it remains unknown if the resulting tight coupling can make poisoning effect disappear.

Our research results are novel from three perspectives. *First*, we provide the theoretical foundation of gradient coupling in poisoning attacks and named it  $\lambda$ -Coupling. We show that the distribution of benign gradients from honest clients can be separable from the distribution of poisoned gradients from compromised clients. *Second*, we analyze the behavior of the  $\lambda$ -Coupling under four scenarios: adaptive attacker, skewed client selection, Non-IID data distribution, and different gradient window sizes. We identify situations where the  $\lambda$ -Coupling would fail as the poisoning detection indicator. *Last*, we investigate how poisoning coupling interacts with two existing defense approaches against poisoning attacks: differential privacy noise and Byzantine-robust aggregation. We show that both methods would create tight coupling, but such gradient behavior does not necessarily neutralize the poisoning effect. Meanwhile, both mitigation approaches could severely deteriorate the benign performance of the trained global model. Our formal analysis and empirical evaluation on three benchmarking datasets validate our understanding of (1) why the eigenvalues of the covariance of the gradient update from the benign clients and the poisoned clients can be separated, (2) under what conditions the  $\lambda$ -Coupling can serve as the forensic evidence for poisoning detection, and (3) how the coupling effect behaves in existing defenses. As federated learning systems become more popular with promises of increased accuracy and privacy, highlighting and understanding these behaviors is an integral part of the poisoning attack mitigation effort.

## 2. Poisoning Attack Threat Model

Poisoning attack in federated learning assumes the existence of an adversary on the compromised client and occurs during the training phase. The attack goal is to change the behavior of the trained global model. The attack can be performed on data or model. Data poisoning has two types: 1) clean-label and 2) dirty-label. Clean-label attacks [21] inject training examples that are cleanly labeled by a certified authority. Imperceptible adversarial watermarks are injected into the clean input to form a poisoning instance with a clean label and simultaneously minimize the distance of the input to the target instance.

In contrast, dirty-label poisoning deletes, inserts or replaces training examples with the desired target label into the training set. One example is the backdoor poisoning [1, 6, 25, 28], in which the adversary inserts small regions of the original training data and modifies the label as the desired target class to embed the trigger

into the model. Accordingly, the unaltered input will not be affected, and the input with the trigger will behave according to the adversary’s objective. Another example is the label-flipping attack [23], which modifies the label of objects from a specific class (attack victim) to another designated target class. The features of the data are kept unchanged. Model poisoning attack happens during the local model training process, by modifying the objective of poison local model updates [8]. Since data poisoning attacks eventually change a subset of updates sent to the model at any given round, model poisoning is believed to subsume data poisoning in federated learning settings [2]. Given that the goal of this paper is to study the coupling effect, which can be observed in both backdoor [24] and label-flipping [7], we focus on the targeted dirty-label poisoning and consider the commonly used label-flipping attack [23].

We make the following assumptions in our threat model. Each malicious client can only manipulate the training data  $X_i$  with auxiliary information, such as the target label on their own device but cannot access or manipulate other participants’ data. The attack corrupts training data with label change, but the learning procedure remains unaltered, e.g., SGD, loss function, or server aggregation. The attack is not specific to any deep learning model architecture, loss function, or optimization function. This attack will only drop the prediction accuracy of the source class. Yet, the poisoning attack has little negative impact on the accuracy for the rest of the classes. Let  $F(x)$  denote the global model trained in federated learning,  $f_i(x)$  be the local model of client  $i$ ,  $(x, y)$  denote the raw data and its ground truth label in the training set of client  $i$ . The attack method  $\rho$  replaces the ground truth label  $y$  to  $y'$  to mislead the joint training so that the federated global model produced by federated learning will be fooled and mispredict examples of source class  $y$  to target class  $y'$  with high confidence, formally:

$$\rho : \rho(x, y) = (x, y')$$

$$s.t. \quad f_i(x) = y', \quad y' \neq y, \quad \max[F(x) = y']$$

The objective is to maximize the chance of the global model  $F(x)$  misclassifying test examples from the source class into the target class.

The threat model also assumes that only the training data of the source class on a small percentage ( $\lambda$ ) of compromised clients is poisoned. While  $\lambda$  can be small, such as 5% or 10% of total  $N$  clients, the availability of malicious clients can be purposely increased. We follow [23] with  $\alpha\%$  chance that the gradient update collected by the server is from a malicious client. Unless otherwise specified, we adopt  $\alpha = 0.6$  at each round for effective poisoning attacks.

### 3. $\lambda$ -Coupling Effect of Poisoning Attacks

In vanilla federated learning, the eigenvalues of the covariance of the gradient update from the set of training examples for a given class would couple with each other and form only one cluster under PCA. **Figure 1** demonstrates the coupling effect of two classes in CIFAR10 under vanilla federated learning. When the system is under the targeted dirty-label poisoning attack, it is observed that by a small  $\lambda$  malicious clients, the eigenvalues of the covariance of the gradient update from the set of training examples for the source class will consist of two sub-populations: the poisoned gradient on malicious clients and the benign gradient on honest clients. We refer to such phenomenon as  $\lambda$ -Coupling, implying a weak coupling relation between the poisoned and benign gradients.

**Definition 1.  $\lambda$ -Coupling.** Given  $0 < \lambda < 1/2$ , let  $H$  and  $P$  denote the two distributions: Honest and Poisoned, respectively, with finite covariance. Let the mixing loss function be  $\mathbb{G} = (1 - \lambda)H + \lambda P$ ,  $v$  be the top eigenvalue of the covariance of  $\mathbb{G}$  and  $\mu_{\mathbb{G}}$  is the mean of the mixed distribution  $\mathbb{G}$ . The two distributions  $H$  and  $P$  are separable if there exists some  $\tau$  such that:

$$\Pr_{X \sim H} [|\langle X - \mu_{\mathbb{G}}, v \rangle| > \tau] < \lambda, \quad (1)$$

$$\Pr_{X \sim P} [|\langle X - \mu_{\mathbb{G}}, v \rangle| < \tau] < \lambda. \quad (2)$$

From the definition, we can see that when  $0 < \lambda < 1/2$ , the distribution of the benign gradients and the poisoned gradients are loosely coupled. By projecting the high-dimension gradients onto the two-dimension space with principal component analysis (PCA), we illustrate  $\lambda$ -Coupling in **Figure 2** in a detection window of 5 and 10 rounds. Both cases can clearly separate the poisoned gradients (blue cross) on the source class from the benign gradient update from honest clients (yellow dot). These results also show that the  $\lambda$ -Coupling phenomenon persists with different percentages of malicious clients and window sizes as long as  $\lambda$  satisfies  $0 < \lambda < 1/2$ . Therefore, it is possible to inspect the gradient distributions across all clients for outlier detection and removal.

This definition also implies the effect of  $\lambda$ -Coupling in Equation 1 and 2 are dependent with  $v$ , the top eigenvalue of the covariance of  $\mathbb{G}$  and  $\mu_{\mathbb{G}}$ , the mean of the mixed distribution  $\mathbb{G}$ . Since  $\mathbb{G}$  indicates the mixed distribution of honest and poisoned gradients, the distributional differences can be formulated by the mean difference between the two distributions. Therefore, we can have the following Theorem.

**Theorem 1.** Given  $0 < \lambda < 1/2$ , let  $H$  and  $P$  denote Honest and Poisoned distributions, with mean  $\mu_H, \mu_P$ ,

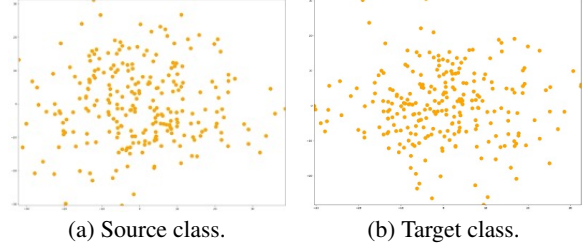


Figure 1:  $\lambda$ -Coupling for CIFAR10 with source class (Car) and target class (Truck) under no poisoning attack.

and finite covariance  $\Sigma_P, \Sigma_H \preceq \phi^2 \mathbb{I}$ . Let the mixing loss function be  $\mathbb{G} = (1 - \lambda)H + \lambda P$  and  $\Delta = \mu_H - \mu_P$ .

Then, if  $\lambda \geq \frac{6\phi^2}{\|\lambda\Delta\|_2^2}$ ,  $P$  and  $H$  satisfy  $\lambda$ -Coupling.

*Proof.* We first prove  $|\langle \Delta, v \rangle| > \frac{2\phi}{\sqrt{\lambda}}$  under the

assumption of  $\|\Delta\|_2^2 \geq \frac{6\phi^2}{\lambda}$ . Given  $\mathbb{G} = (1 - \lambda)H + \lambda P$ , we have  $\mu_{\mathbb{G}} = (1 - \lambda)\mu_H + \lambda\mu_P$  and

$$\mathbb{E}_{X \sim H} [(X - \mu_{\mathbb{G}})(X - \mu_{\mathbb{G}})^T] = \Sigma_H + \lambda^2 \Delta \Delta^T$$

$$\mathbb{E}_{X \sim P} [(X - \mu_{\mathbb{G}})(X - \mu_{\mathbb{G}})^T] = \Sigma_P + (1 - \lambda)^2 \Delta \Delta^T$$

Since  $\mathbb{G}$  is a mixed distribution of  $H$  and  $P$ , we have

$$\begin{aligned} \Sigma_{\mathbb{G}} &= (1 - \lambda) \mathbb{E}_{X \sim H} [(X - \mu_{\mathbb{G}})(X - \mu_{\mathbb{G}})^T] \\ &\quad + \lambda \mathbb{E}_{X \sim P} [(X - \mu_{\mathbb{G}})(X - \mu_{\mathbb{G}})^T] \\ &= (1 - \lambda) \Sigma_H + \lambda \Sigma_P + \lambda(1 - \lambda) \Delta \Delta^T \end{aligned}$$

Since the  $l_2$  norm of the matrix is the largest singular value, we have  $\|\Delta \Delta^T\|_2 = \|\Delta\|_2^2$ . And subsequently:

$$\begin{aligned} \lambda(1 - \lambda) \Delta \Delta^T &= \lambda(1 - \lambda) \|\Delta\|_2^2 \leq \|\Sigma_{\mathbb{G}}\|_2 = v^T \Sigma_{\mathbb{G}} v \\ &= (1 - \lambda) v^T \Sigma_H v + \lambda v^T \Sigma_P v + \lambda(1 - \lambda) \langle \Delta, v \rangle^2 \\ &\leq \phi^2 + \lambda(1 - \lambda) \langle \Delta, v \rangle^2. \end{aligned}$$

The second line is due to  $\Sigma_{\mathbb{G}} \succeq \lambda(1 - \lambda) \Delta \Delta^T$ . So we have  $\|\Sigma_{\mathbb{G}}\|_2 \geq \lambda(1 - \lambda) \|\Delta\|_2^2$ . The third line holds given that the  $l_2$  norm of the matrix equals its top eigenvalue for a symmetric orthogonal matrix. Since by assumption  $\phi^2 \leq \frac{\lambda}{6} \|\Delta\|_2^2$  and  $0 \leq \lambda \leq 1/2$ , we have:

$$\langle \Delta, v \rangle^2 \geq \left(1 - \frac{1}{6(1 - \lambda)}\right) \|\Delta\|_2^2 \geq 2/3 \|\Delta\|_2^2 \geq \frac{4\phi^2}{\lambda}.$$

Next we show given  $|\langle \Delta, v \rangle| > \frac{2\phi}{\sqrt{\lambda}}$ , there exists a  $\tau = \lambda |\langle \Delta, v \rangle| + \frac{\phi}{\sqrt{\lambda}}$  such that:

$$\Pr_{X \sim H} [|\langle X - \mu_{\mathbb{G}}, v \rangle| > \tau] < \lambda,$$

$$\Pr_{X \sim P} [|\langle X - \mu_{\mathbb{G}}, v \rangle| < \tau] < \lambda.$$

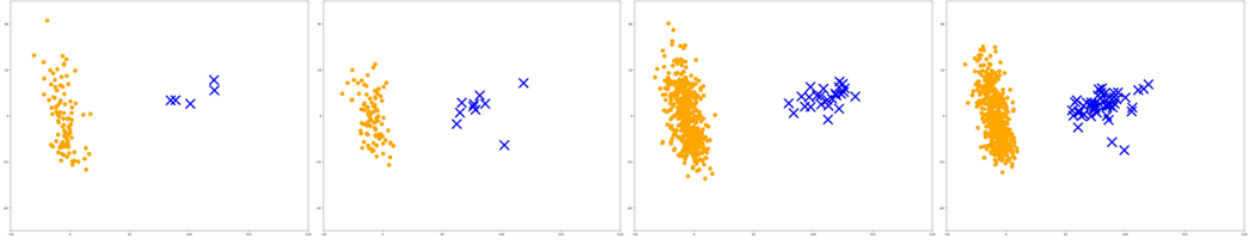


Figure 2:  $\lambda$ -Coupling on CIFAR10 with source (Car) and target (Truck). Yellow dots are benign gradient updates, and blue crosses are poisoned ones.

We first prove the left side. For  $|\langle X - \mu_G, v \rangle| > \tau$ ,

$$\begin{aligned} |\langle X - \mu_H, v \rangle| &= |\langle X - \mu_G, v \rangle - \lambda \langle \Delta, v \rangle| \\ &\geq |\langle X - \mu_G, v \rangle| - \lambda |\langle \Delta, v \rangle| \\ &> \tau - \lambda |\langle \Delta, v \rangle| = \frac{\phi}{\sqrt{\lambda}} \end{aligned}$$

The second line holds due to triangle inequality, and the third line is due to  $|\langle X - \mu_G, v \rangle| > \tau$ . Therefore,

$$\begin{aligned} \Pr_{X \sim H} [|\langle X - \mu_G, v \rangle| > \tau] &\leq \Pr_{X \sim H} [|\langle X - \mu_H, v \rangle| > \frac{\phi}{\sqrt{\lambda}}] \\ &\leq \lambda. \end{aligned}$$

The right-hand side is due to Chebyshev's inequality.

Then we prove the right side:  $|\langle X - \mu_G, v \rangle| < \tau$ ,

$$\begin{aligned} |\langle X - \mu_P, v \rangle| &= |\langle X - \mu_G, v \rangle - (1 - \lambda) \langle \Delta, v \rangle| \\ &\geq (1 - \lambda) |\langle \Delta, v \rangle| - |\langle X - \mu_G, v \rangle| \\ &\geq (1 - \lambda) |\langle \Delta, v \rangle| - \tau \\ &= (1 - \lambda) |\langle \Delta, v \rangle| - \lambda |\langle \Delta, v \rangle| - \frac{\phi}{\sqrt{\lambda}} \\ &= (1 - 2\lambda) |\langle \Delta, v \rangle| - \frac{\phi}{\sqrt{\lambda}} \\ &> (1 - 2\lambda) \frac{2\phi}{\sqrt{\lambda}} - \frac{\phi}{\sqrt{\lambda}} \\ &= \frac{\phi}{\sqrt{\lambda}} - 4\sqrt{\lambda}\phi > \frac{\phi}{\sqrt{\lambda}}. \end{aligned}$$

The second line is due to triangle inequality, and the third line is because  $|\langle X - \mu_G, v \rangle| < \tau$ . The fourth line due to the assumption  $\tau = \lambda |\langle \Delta, v \rangle| + \frac{\phi}{\sqrt{\lambda}}$  and line six due to the assumption  $|\langle \Delta, v \rangle| > \frac{2\phi}{\sqrt{\lambda}}$ . Therefore,

$$\begin{aligned} \Pr_{X \sim P} [|\langle X - \mu_G, v \rangle| > \tau] &\leq \Pr_{X \sim P} [|\langle X - \mu_P, v \rangle| > \frac{\phi}{\sqrt{\lambda}}] \\ &\leq \lambda. \end{aligned}$$

The right-hand side holds due to Chebyshev's inequality. Thus completes the proof.  $\square$

	Fashion-MNIST	CIFAR10	LFW
# training data	60000	50000	2267
# validation data	10000	10000	756
# features	28*28	32*32*3	32*32*3
# classes	10	10	62
# data/client	600	500	300
# local iteration $L$	60	50	100
local batch size $B$	10	10	3
# rounds $T$	100	200	60
accuracy	0.893	0.743	0.695

Table 1: Benchmark datasets and parameters

Theorem 1 implies that the gradient updates collected at the server would demonstrate two separated clusters when the federated learning is under poisoning attack, if (1)  $0 < \lambda < 1/2$  and (2)  $\lambda \geq \frac{6\phi^2}{\|\lambda\Delta\|_2^2}$ . Given

that the honest and poisoned distributions are different under different datasets, data instances, and models, the mean  $\mu_H, \mu_P$  and finite covariance  $\Sigma_P, \Sigma_H \preceq \phi^2 \mathbb{I}$  are different as well. Therefore, the effect of  $\lambda$ -Coupling is data-dependent and model-dependent. By rewriting the second item as  $\|\Delta\|_2^2 \geq \frac{6\phi^2}{\lambda}$ , we can find the minimum distribution difference that would enable  $\lambda$ -Coupling under a given  $\lambda$ .

## 4. Evaluating $\lambda$ -Coupling Effect

To investigate the behavior of  $\lambda$ -Coupling in poisoning attacks, we conduct the study on three benchmark datasets. Fashion-MNIST [27] is a grey-scale image dataset associated with 10 clothing classes such as T-shirt, Trouser, and Pullover. CIFAR10 [14] is a dataset of colored images from 10 object classes such as dog, car, and plane. LFW [13] is a human face dataset. Since most classes have a very limited number of data points, we consider 3023 images from 62 classes with more than 20 images per class. **Table 1** provides a detailed configuration of these datasets. For the label-flipping poisoning attacks, we consider source-target pairs Trouser  $\Rightarrow$  Ankle boot for Fashion-MNIST, Car  $\Rightarrow$  Truck for CIFAR10, Angelina Jolie  $\Rightarrow$  Jennifer Aniston for LFW, respectively. Results on other source and target classes demonstrate a similar











sample	target	$\lambda$	before outlier removal		after outlier removal		sample	target	m	before outlier removal		after outlier removal		
			victim class	rest classes	victim class	rest classes				victim class	rest classes	victim class	rest classes	
	T-shirt	benign	92.3%	88.9%	92.3%	88.9%		ankle	5%	benign	97.0%	88.4%	96.9%	88.4%
		5%	77.0%	88.6%	92.2%	88.5%					82.2%	88.3%	96.9%	88.4%
		10%	41.6%	88.6%	92.2%	88.4%					44.9%	88.1%	96.7%	88.2%
	sneaker	benign	84.2%	89.9%	84.2%	89.9%		shirt	5%	benign	92.5%	88.9%	92.5%	88.9%
		5%	71.7%	89.7%	84.0%	89.6%					76.0%	88.6%	92.4%	88.8%
		10%	38.6%	89.4%	84.0%	89.3%					51.6%	88.5%	92.3%	88.7%
	horse	benign	73.3%	74.4%	73.3%	74.4%		truck	5%	benign	88.1%	72.6%	88.1%	72.6%
		5%	59.1%	74.4%	73.1%	74.3%					75.6%	72.1%	87.9%	72.4%
		10%	48.3%	74.3%	73.0%	73.8%					50.3%	72.0%	87.5%	72.3%
	bird	benign	79.7%	73.7%	79.7%	73.7%		cat	5%	benign	78.4%	73.8%	78.4%	73.7%
		5%	64.5%	73.7%	79.6%	73.5%					66.5%	73.3%	78.1%	73.4%
		10%	48.9%	73.7%	79.4%	73.3%					40.3%	73.5%	79.7%	73.3%
	Tiger Woods	benign	70.6%	69.4%	70.4%	69.1%		Jennifer Aniston	5%	benign	68.7%	69.6%	68.4%	69.6%
		5%	62.3%	68.7%	70.2%	68.5%					59.1%	69.2%	68.4%	69.2%
		10%	51.1%	69.3%	70.1%	68.3%					46.8%	68.9%	67.8%	69.0%

Figure 3: Micro f1 score before and after the poisoning detection and removal.

phenomenon. To maximize the poisoning effect, we follow [23] and start to inject malicious clients 10 rounds before the end of training for Fashion-MNIST and 40 rounds for CIFAR10 and LFW. We use  $\lambda = 10\%$  and  $\alpha = 0.6$  unless otherwise specified. Besides PCA to capture the two cluster phenomenon, we apply KMeans with the classic Lloyd algorithm and set the number of clusters to 2 with 10 initialization seeds and maximum iteration to 300. The relative tolerance regarding the Frobenius norm of the difference in the cluster centers of two consecutive iterations to declare convergence is set to 0.0004.

We measure the class-wise results using micro f1 score, which is the harmonic mean of Precision and Recall:  $f1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$ . Precision  $\frac{tp}{tp+fp}$  is the ratio of correctly identified positive samples in all predicted positive samples, where  $tp$  is the number of true positives, and  $fp$  is the number of false positives. Recall  $\frac{tp}{tp+fn}$  is the ratio of correctly identified positive samples in all observed positive instances, where  $fn$  is the number of false negatives. Our federated learning setup follows the simulator in [23] with a total of  $N = 100$  clients and the number of participating clients  $K_t = 10\%$  of  $N$  in each round.

#### 4.1. Vanilla $\lambda$ -Coupling Evaluation

Recall Theorem 1, the distribution of benign gradients from honest clients is separable from that of poisoned gradients when the percentage  $\lambda$  of compromised clients satisfies  $0 < \lambda < 1/2$ . **Figure 3** reports the results under the scenario with no poisoning and poisoning scenarios with 5% and 10% attackers. We make two observations: (1) The poisoning attacks under both  $\lambda$  settings are stealthy. They succeed in dropping the prediction accuracy of the source

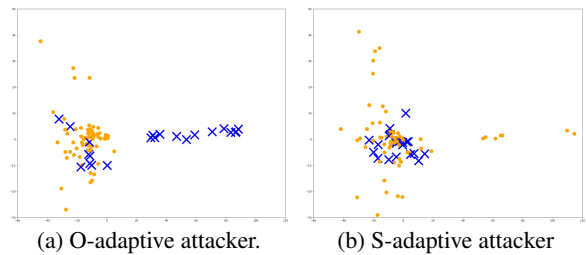


Figure 4:  $\lambda$ -Coupling under adaptive attackers on Fashion-MNIST. Source (Trouser)  $\Rightarrow$  target (Ankle boot).

class by up to 24% and yet have a little negative impact on the prediction for the rest of the classes. (2) With  $\lambda$ -Coupling-based poisoning detection and removal, the defense can safeguard federated learning systems against targeted poisoning attacks. While we provide the theoretical cause of the  $\lambda$ -Coupling, existing defenses against the targeted poisoning attacks are mainly built upon the empirical observation [12, 22, 24, 26] of gradient separation with different concrete implementation techniques.

However, simply relying on analyzing the eigenvalue of the client's gradient update over one or multiple rounds to conclude that the smaller cluster contains the poisoned gradient updates and expel those corresponding clients may not be accurate.

#### 4.2. No Guaranteed Existence of Two Clusters

We first examine whether there must be two separate clusters for the targeted dirty-label poisoning attacks. We argue that the  $\lambda$ -Coupling is not guaranteed.

**O-adaptive attacker.** The first scenario is the Occasional adaptive attacker who aims to bypass anomaly detection. For instance, the malicious clients camouflage in the crowd of the benign clients and only perform the label flipping occasionally instead of each

		vanilla poisoning	O-adaptive		S-adaptive		Noise injection	
			10%	20%	60%	80%	$\mathcal{N}(0, 0.05^2)$	$\mathcal{N}(0, 1^2)$
Fashion-MNIST	no poisoning	0.97	0.97	0.97	0.97	0.97	0.964	0.913
	$\lambda=5\%$	0.822	0.834	0.866	0.871	0.845	0.851	0.805
	$\lambda=10\%$	0.449	0.523	0.581	0.602	0.517	0.823	0.726
CIFAR10	no poisoning	0.881	0.881	0.881	0.881	0.881	0.868	0.809
	$\lambda=5\%$	0.756	0.771	0.795	0.794	0.758	0.746	0.677
	$\lambda=10\%$	0.503	0.511	0.534	0.582	0.516	0.732	0.618
LFW	no poisoning	0.687	0.687	0.687	0.687	0.687	0.681	0.644
	$\lambda=5\%$	0.591	0.616	0.644	0.625	0.592	0.577	0.565
	$\lambda=10\%$	0.468	0.47	0.492	0.561	0.483	0.465	0.556

Table 2: F1-score of the victim source class under adaptive poisoning attacker and Gaussian noise. For the poisoning source class and target class, Trouser  $\Rightarrow$  Ankle boot for Fashion-MNIST, Car  $\Rightarrow$  Truck for CIFAR10, Angelina Jolie  $\Rightarrow$  Jennifer Aniston for LFW, respectively.

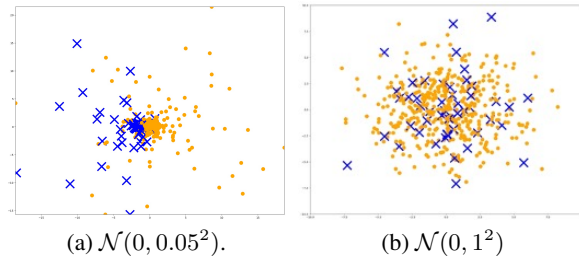


Figure 5:  $\lambda$ -Coupling under Gaussian noise on Fashion-MNIST. Source (Trouser)  $\Rightarrow$  target (Ankle boot).

time it is selected. **Figure 4a** demonstrate the PCA result of the client gradients in 5 rounds (round 45 - 50) under the O-adaptive attacker. The malicious clients, even though being selected to participate in one round, choose not to poison the local dataset during the local model training at the percentage of 10%. When O-adaptive attackers contribute to the cluster with poisoned and benign gradient updates, they can circumvent the  $\lambda$ -Coupling-based detection. **Table 2** indicates that the attack effect of the occasional adaptive attack would be weaker compared to the vanilla poisoning attack due to the fewer chances of the poisoned gradient update.

**S-adaptive attacker.** The second scenario is the Selective adaptive attacker. The malicious clients only poison a selected percentage of the training examples in its datasets. The corresponding local gradient updates are computed over a mixture of clean and poisoned training samples. **Figure 4b** demonstrate the PCA result of the client gradients in 5 rounds (round 45 - 50) of the S-adaptive attacker. Even though the malicious clients are selected to participate in one round, they can choose only to flip 60% of the victim source class they own. Consequently, the  $\lambda$ -Coupling cannot have two cleanly separated clusters under the S-adaptive attacker. Table 2 shows the effect of the selective adaptive attacker. The gradient update mixed with benign and poisoned data is less destructive on the poisoning attack effect when compared to vanilla poisoning.

**Impact of noise.** In addition to the O-adaptive and

the S-adaptive attacker, we find that injecting a large amount of noise would also blend the two clusters. The resulting tight coupling is because noise injection takes a uniform effort on both the benign gradients from the honest clients and the poisoned gradients from the malicious clients. Therefore, the two clusters under the  $\lambda$ -Coupling could either come close to one another as shown in **Figure 5a** or even mix together when the noise is large enough as shown in **Figure 5b**. Table 2 shows the effect of Gaussian noise with zeros means and variance  $0.05^2$  and  $1^2$ . We can see that the injected noise would lower the f1 score performance even without the poisoning attack. When  $\lambda$  is 5% or 10%, a small noise  $\mathcal{N}(0, 0.01^2)$  does not affect the poisoning effect much. However, a large noise would compensate for the poisoning effect on the victim but at the cost of a larger f1 score drop on the rest of the classes.

### 4.3. Smaller Clusters Must Be Malicious? No.

Next, we explore whether the smaller cluster in the  $\lambda$ -Coupling always represents the gradient update from the malicious clients. We consider three scenarios: attacker domination with a large  $\lambda$ , skillful attacker domination with a small  $\lambda$  but a large  $\alpha$ , and non-IID federated learning. We argue that the smaller cluster may not always reflect the malicious clients.

**Poisoning Attacker Domination.** Due to the distributed nature of federated learning, the server has no control over the client. The malicious clients may take over the gradient direction. **Figure 6a** demonstrates the  $\lambda$ -Coupling in one round at round 50 when  $\lambda = 60\%$ . Even when the malicious clients are not purposely available (no  $\alpha$  control), the gradient update with random sampling is dominated by the poisoned gradient, and the larger cluster represents the malicious client.

**Skillful Poisoning Attacker Domination.** As discussed earlier, the adversary could purposely increase its availability to cause damage to the federated learning model. Therefore, even though the percentage  $\lambda$  of the malicious client is small, the attacker could take

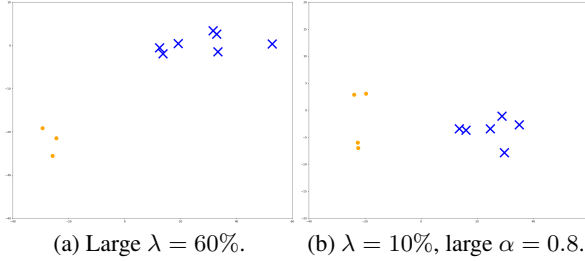


Figure 6:  $\lambda$ -Coupling under different data distribution on Fashion-MNIST. Source (Trouser)  $\Rightarrow$  target (Ankle boot).

control of the gradient update on the victim source class. **Figure 6b** shows the  $\lambda$ -Coupling in one round of federated learning (round 50) when  $\lambda = 10\%$  but  $\alpha = 0.8$ . The poisoned gradient update out-numbers the benign gradient update, and the larger cluster represents the malicious client. As such, the randomness in client selection can lead to an incorrect revocation.

**Non-IID Setting.** In the non-IID setting, the benign gradients may construct two different distributions. For instance, when we have 30 out of 100 clients who do not have class Trouser in their local dataset, the gradient update from the 70 clients with class Trouser and the gradient update from the 30 clients without class Trouser would form two clusters. **Figure 7a** visualize the  $\lambda$ -Coupling in five rounds of federated learning (round 45- 50) caused by the vanilla non-IID data distribution. We argue that the  $\lambda$ -Coupling can cause false positive detection of the poisoned gradients in the non-IID setting without the attacker.

Next, we inject poisoning attacks into the non-IID setting. Specifically, among 70 out of 100 clients,  $\lambda = 10\%$  and  $\alpha = 0.6$  is applied. **Figure 7b** visualize the  $\lambda$ -Coupling in five rounds of federated learning (round 45- 50) under the non-IID data distribution with poisoning. We observe that there are three clusters under PCA, making it even harder to determine which cluster represents the poisoning attacker. Given the typical use case of the non-IID setting in federated learning, our result indicates that existing defense approaches leveraging the  $\lambda$ -Coupling may fail as an indicator for poisoning attack mitigation.

#### 4.4. The Larger Detection Window No Better

Existence of the Occasional adaptive attacker makes a per-round based  $\lambda$ -Coupling analysis inaccurate, and the randomness in federated learning client selection may result in different portions of the malicious client. Both may lead to incorrect outlier detection and removal since the smaller cluster may not be guaranteed to be the poisoned gradient update. Thus, the one-round-based detection window may not be ideal for  $\lambda$ -Coupling

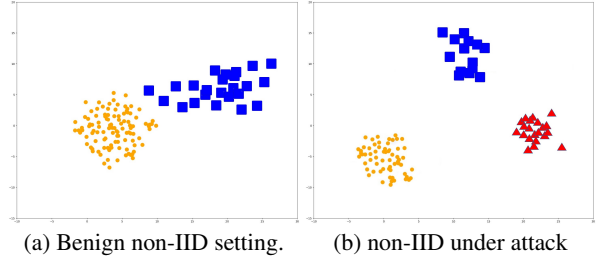


Figure 7:  $\lambda$ -Coupling under non-IID setting. Yellow dots mean benign clients with the source class (Trouser), Blue squares represent benign clients without the source class, and red triangles denote malicious clients.

analysis. We demonstrate the detection window size of 1 (round 50), 5 (round 45-50), 25 (round 25-50), and 50 (round 1-50) under for  $\lambda$ -Coupling analysis using PCA in **Figure 8** and KMeans in **Figure 9**. We show that it is prone to have false positives on cluster analysis when the detection window is large, and such false positive instances persist for both PCA and KMeans-based  $\lambda$ -Coupling analysis. The two clusters are not as cleanly separated when the detection window is 25 rounds or 50 rounds when compared to 1 round and 5 rounds. Therefore, neither a large nor small detection window is guaranteed to capture the  $\lambda$ -Coupling well.

## 5. $\lambda$ -Coupling in Poisoning Mitigation

In this section, we study the coupling effect of poisoning attacks under two existing poisoning mitigation strategies: differential privacy controlled noise [1, 15, 19, 25] and byzantine-robust aggregation [1, 8]. We resolve the misunderstanding that the tight coupling of benign and poisoned gradients brought by these approaches does not necessarily make the poisoning effect disappear.

**$\lambda$ -Coupling under Differential Privacy.** Differentially private federated learning [10, 17] uses privacy parameter controlled noise to perturb the gradients before performing server-side stochastic gradient descent for participation-level differential privacy guarantee. While the technique is not designed to defend against poisoning attacks, the two key steps effectively limit the poisoning effect. First, the local parameter update shared for the differentially private global aggregation is clipped to bound its sensitivity  $S$  in terms of its  $l_2$  norm. Second, Gaussian noise  $\mathcal{N}(0, \sigma^2 S^2)$  is added to the local parameter update for sanitization, where  $\sigma$  is defined as the noise scale.

**Table 3** shows the results of poisoning attacks under different kinds of differential privacy noise parameter settings. We make three observations: (1) a small differential privacy noise would provide insufficient protection against poisoning attacks. Under  $\lambda = 5\%$

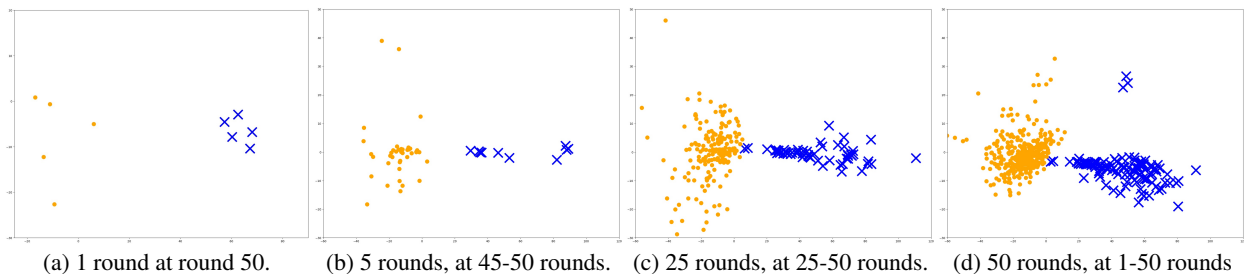


Figure 8: Comparisons of  $\lambda$ -Coupling using PCA, with different detection window sizes.

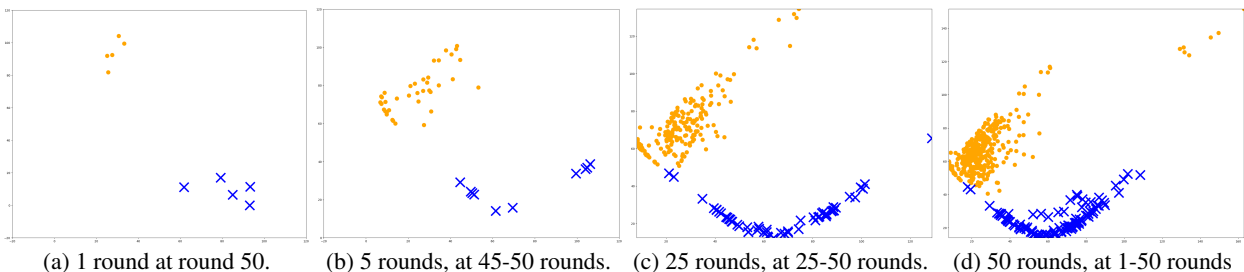


Figure 9: Comparisons of  $\lambda$ -Coupling using KMeans, with different detection window sizes.

		$\lambda$	no attack	1%	5%	10%
Fashion-MNIST	vanilla	victim class	0.97	0.948	0.822	0.449
		rest classes	0.884	0.883	0.883	0.881
	$\sigma=0.1, S=0.1$	victim class	0.956	0.931	0.799	0.451
		rest classes	0.876	0.876	0.877	0.876
	$\sigma=2, S=0.5$	victim class	0.856	0.847	0.809	0.754
		rest classes	0.782	0.782	0.782	0.782
$\sigma=0.5, S=0.1$	victim class	0.883	0.881	0.86	0.829	
	rest classes	0.834	0.834	0.834	0.833	
CIFAR10	vanilla	victim class	0.881	0.856	0.756	0.503
		rest classes	0.726	0.726	0.721	0.720
	$\sigma=0.1, S=0.1$	victim class	0.855	0.821	0.704	0.506
		rest classes	0.722	0.723	0.723	0.723
	$\sigma=2, S=0.5$	victim class	0.725	0.719	0.692	0.63
		rest classes	0.609	0.608	0.608	0.609
$\sigma=0.4, S=0.1$	victim class	0.787	0.783	0.725	0.701	
	rest classes	0.681	0.682	0.682	0.683	
LFW	vanilla	victim class	0.687	0.667	0.591	0.468
		rest classes	0.696	0.696	0.692	0.689
	$\sigma=0.2, S=0.1$	victim class	0.669	0.64	0.589	0.473
		rest classes	0.679	0.679	0.68	0.679
	$\sigma=2, S=0.5$	victim class	0.614	0.611	0.581	0.503
		rest classes	0.637	0.637	0.637	0.637
$\sigma=0.8, S=0.1$	victim class	0.654	0.649	0.602	0.578	
	rest classes	0.662	0.662	0.661	0.661	

Table 3: Impact of differential privacy noise on the targeted dirty-label poisoning, measured in Micro f1.

malicious clients, poisoning attacks may cause about 10% to 18% accuracy loss on the source class being poisoned, and when  $\lambda = 10\%$ , the poisoning effect may cause 21% to 52% accuracy lost on the victim source class. Recall Section 4.2, we can observe  $\lambda$ -Coupling when the injected noise is small but the poisoning effect remains strong. By comparison, a large noise would offer good protection but at the cost of accuracy. For example, the f1 score of the rest classes other than the victim class drops 10.2% on Fashion-MNIST, 11.7% on CIFAR10, and 5.9% on LFW. (2) Empirically, we find the differential privacy noise setting that best balance poisoning protection and overall accuracy utility:  $\sigma=0.5, S=0.1$  for Fashion-MNIST,  $\sigma=0.4, S=0.1$  for CIFAR10,

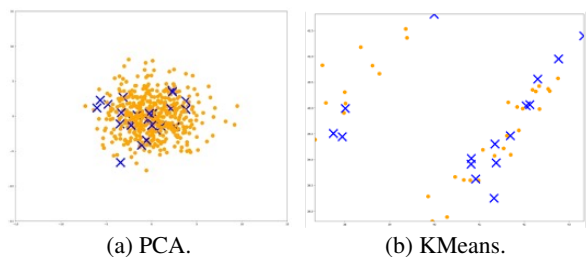


Figure 10:  $\lambda$ -Coupling under sufficient differential privacy noise.

$\sigma=0.8, S=0.1$  for LFW. However, it has been challenging to search for the right balance of noise to bring back the gradient from poisoning. Yet the noise injection is not too much to prevent federated learning from converging to a reasonable accuracy point. This is because such privacy noise setting is dataset-dependent and model-dependent. As shown in **Figure 10**, tight coupling is observed, and the malicious gradients are blended with the benign gradients under PCA and KMeans. While differential privacy noise encourages tight coupling, the poisoning effect can be partially neutralized but does not go away. (3) the additional clipping operation in the differentially private training of noise setting  $\sigma=2, S=0.5$  makes the model slightly better protected against the poisoning attacks when compared to the random noise alone (recall Table 2).

While adding noise delays the poisoning effect, we show that such mitigation is only helpful when the number of adversaries is small. The perturbed gradient must be an  $e^\epsilon - 1$  dominating strategy [18] slightly deviated from the main-stream gradients to be



		$\lambda$	no attack	20%	40%	60%
F-MNIST	vanilla	victim class	0.97	0.172	0	0
		rest classes	0.884	0.882	0.882	0.882
	$\sigma=0.5, S=0.1$	victim class	<b>0.883</b>	<b>0.536</b>	<b>0.149</b>	<b>0</b>
		rest classes	0.834	0.829	0.828	0.831
CIFAR-10	vanilla	victim class	0.881	0.104	0	0
		rest classes	0.726	0.723	0.726	0.722
	$\sigma=0.4, S=0.1$	victim class	<b>0.787</b>	<b>0.441</b>	<b>0.112</b>	<b>0</b>
		rest classes	0.681	0.678	0.684	0.677
LFW	vanilla	victim class	0.687	0.035	0	0
		rest classes	0.696	0.695	0.693	0.694
	$\sigma=0.8, S=0.1$	victim class	<b>0.654</b>	<b>0.371</b>	<b>0.076</b>	<b>0</b>
		rest classes	0.662	0.659	0.655	0.658

Table 4: Impact of differential privacy noise in Micro f1 when the malicious clients dominate the training,  $\alpha = 0.6$ .

protective. As shown in **Table 4**, when we set  $\lambda = 60\%$ , the majority of the gradient updates on the source victim class at each round would optimize towards the poisoning target class. Then, differential privacy noise would preserve the poisoning dominance.

**$\lambda$ -Coupling under Byzantine-Tolerant Aggregation.** Recent proposals for Byzantine-tolerant distributed learning are also frequently studied in the context of poisoning mitigation. We evaluate  $\lambda$ -Coupling with a representative approach called Krum [3]. Krum selects  $\kappa$  number of models that is most similar to other models as the global models for the next round by computing pairwise distances between all models submitted in a given round and summing up the  $K_t - \kappa - 2$  closest distances for each model. We consider the  $\kappa$  Byzantine participants the same number as  $\lambda$ , which is the percentage of malicious clients in training. We present the results in **Table 5** and make four observations. (1) Krum can mitigate the targeted poisoning attack when the percentage  $\lambda$  of malicious clients is small. (2) When  $\lambda$  is as large as 20% or 40%, models most similar to other models contain both benign and malicious gradients. Thus, even with tight coupling, the poisoning effect does not go away. (3) Krum causes significant degradation in the performance of the global model, even in the absence of attacks. (4) Krum bring significant additional costs to federated learning. The computation cost of finding the most similar model in Krum brings the cost of federated learning up to 3.52s per round for Fashion-MNIST, 8.33s for CIFAR10, and 7.52s for LFW.

## 6. Related Work and Contributions

Existing defense solutions against targeted dirty-label poisoning assume that the federated learning server is trusted and can detect anomalies by separating poisoned contributions from non-poisoned contributions. The defender can flag those poisoned gradients and remove them. [23] directly apply PCA on the local model updated collected over multiple rounds. [24] perform spectral analysis with SVD to generate two clusters. [22] identify the indicative

		$\lambda$	no attack	5%	10%	20%	40%	cost
F-MNIST	vanilla	victim	0.97	0.822	0.449	0.172	0	1.96s
		rest	0.884	0.883	0.881	0.882	0.882	
	Krum	victim	0.925	0.903	0.874	0.669	0.105	3.52s
		rest	0.813	0.813	0.813	0.812	0.812	
CIFAR-10	vanilla	victim	0.881	0.756	0.503	0.104	0	3.83s
		rest	0.726	0.721	0.720	0.723	0.726	
	Krum	victim	0.801	0.786	0.746	0.503	0.097	8.33s
		rest	0.66	0.661	0.662	0.660	0.661	
LFW	vanilla	victim	0.687	0.591	0.468	0.035	0	3.06s
		rest	0.696	0.692	0.689	0.695	0.693	
	Krum	victim	0.587	0.573	0.551	0.433	0.075	7.52s
		rest	0.646	0.645	0.643	0.642	0.640	

Table 5: Impact of byzantine-tolerant aggregation in Micro f1. Cost measured by second per round.

Contribution 1	provided theoretical foundation of the weak coupling effect in data poisoning
Contribution 2	showed weak coupling effect could fail as the attack indicator
Contribution 3	demonstrated tight coupling brought by poisoning mitigation approaches may not neutralize the poisoning effect

Table 6: Contributions and Takeaways

features for comparison by collecting masked user features. These approaches are based on the theoretical foundation studied in Section 3 but with different representations on the eigenvalues of the covariance of the gradient update [12, 26]. Another line of work concerns differential privacy controlled noise injection [1, 15, 19, 25]. However, it is challenging to find the right amount of noise that can largely mitigate the poisoning effect while maintaining the good accuracy of the main task. Byzantine-robust aggregation rules [3, 9, 11, 29] are another option for targeted poisoning mitigation. [5] claims their approach is robust against label-flipping attacks. However, they require the service provider to collect a clean small training dataset, which may incur privacy concerns when the server is not trusted.

Our takeaway on the gradient coupling effect of data poisoning attack in federated learning is three-fold. We summarize the contributions in **Table 6**. Our statistical characterization with strong empirical evidence provides transformative enlightenment on mitigation strategies towards effective countermeasures against present and future data poisoning attacks in federated learning.

## 7. Conclusion

We have presented the first study on the gradient coupling effect of poisoning attacks in federated learning. We formulated the  $\lambda$ -Coupling phenomenon of poisoning attacks through formal analysis and empirical evidence from extensive experimentation. We showed that the distribution of benign gradients from honest clients could be separable from the distribution of poisoned gradients from compromised

clients. Then we analyzed the four failure situations where the  $\lambda$ -Coupling fail to serve as the poisoning detection indicator: adaptive attacker, skewed client selection, non-IID data distribution, and different gradient window sizes. At last, we studied the behavior of  $\lambda$ -Coupling under two poisoning mitigation approaches: differential privacy-controlled noise and byzantine-robust aggregation. We observed that the resulting tight coupling under these defenses cannot neutralize the poisoning effect. The two defense methods also severely lower the performance of the trained global model while at additional communication and computation costs.

**Acknowledgement.** This research is partially sponsored by the NSF CISE grants 2302720, 2312758, 2038029, an IBM faculty award, and a grant from CISCO Edge AI program.

## References

- [1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *AISTATS*, 2020.
- [2] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In *ICML*, 2019.
- [3] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *NeurIPS*, 2017.
- [4] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *MLSys*, 2019.
- [5] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. FTrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*, 2022.
- [6] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.
- [7] Ka-Ho Chow, Ling Liu, Wenqi Wei, Fatih Ilhan, and Yanzhao Wu. StdLens: Model hijacking-resilient federated learning for object detection. In *IEEE/CVF CVPR*, 2023.
- [8] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX Secur.*, 2020.
- [9] Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. The limitations of federated learning in sybil settings. In *USENIX RAID*, 2020.
- [10] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.
- [11] Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *ICML*, 2018.
- [12] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: defending against backdoor attacks using robust statistics. In *ICML*, 2021.
- [13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical report*, 2008.
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. In *Technical report*, 2009.
- [15] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. In *IJCAI*, 2019.
- [16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.
- [17] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *ICLR*, 2018.
- [18] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symp. Found. Comput. Sci.*, 2007.
- [19] Thien Duc Nguyen, Phillip Rieger, Roberta De Viti, Huili Chen, Björn B Brandenburg, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, et al. Flame: Taming backdoors in federated learning. In *USENIX Secur.*, 2022.
- [20] Bjarne Pfitzner, Nico Steckhan, and Bert Arnrich. Federated learning in a medical context: A systematic literature review. *ACM Trans. Internet Technol.*, 21(2):1–31, 2021.
- [21] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suci, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.
- [22] Shiqi Shen, Shruti Tople, and Prateek Saxena. Auror: Defending against poisoning attacks in collaborative deep learning systems. In *ACSAC*, 2016.
- [23] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems. In *ESORICS*, 2020.
- [24] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018.
- [25] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *NeurIPS*, 2020.
- [26] Yunjuan Wang, Poorya Mianjy, and Raman Arora. Robust learning for data poisoning attacks. In *ICML*, 2021.
- [27] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
- [28] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *ICLR*, 2019.
- [29] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, 2018.